CHAPITRE 1
NOTIONS FONDAMENTALES

1. LE SIGNE LINGUISTIQUE
2. LE MOT ET LE MORPHÈME
3. LE LEXIQUE ET LA GRAMMAIRE
4. LE LEXIQUE ET SES USAGES
5. LE FRANÇAIS FONDAMENTAL ET LA STATISTIQUE LEXICALE
6. LE LEXIQUE ET L'UNIVERS

Avant d'aborder l'étude morphologique et sémantique du lexique, il nous semble indispensable de définir les notions fondamentales utilisées en lexicologie. Il nous paraît également important de proposer notre terminologie. En effet, les discussions suscitées entre linguistes par la définition de ces notions ne vont pas sans une grande variété dans la terminologie. Nous essaierons de définir aussi clairement que possible celle que nous emploierons et, dans une perspective pédagogique, d'utiliser la plus simple possible.

Au cours de ce chapitre, on mettra d'abord en lumière la nature du signe linguistique dans l'univers des signes. On s'intéressera ensuite aux différents types de mots qui constituent le lexique. On verra que parmi les unités lexicales du français, il y a des mots simples et des mots construits qui entretiennent des relations paradigmatiques et syntagmatiques avec d'autres mots. On s'intéressera aussi aux différences qui existent entre le lexique et la grammaire, ainsi qu'aux usages du lexique et à la notion de variation. Le volet suivant s'attachera au français fondamental et à la statistique lexicale. À la fin de cette partie consacrée à la présentation des notions fondamentales, l'on examinera les rapports qui existent entre le lexique et l'univers.

1. LE SIGNE LINGUISTIQUE

1.1. L'UNIVERS DES SIGNES

Le signe linguistique appartient à l'univers des signes. Il est important de distinguer d'abord le signe linguistique des autres signes. Un **signe**, au sens le plus large, désigne un élément X qui représente un autre élément Y ou lui sert de substitut. Dans le vocabulaire technique de la **sémiologie** (= science qui étudie tous les procédés ou

systèmes de communication et de signification), un signe est une entité composée de deux éléments solidaires : une **forme** et un **sens**. La forme est un élément perceptible par les organes sensoriels, par exemple un tracé que l'on peut voir, ou un son, simple ou complexe, que l'on peut entendre. En fait, il existe de nombreuses tentatives de classification des signes. Ici, on a choisi d'adopter le classement qui propose les distinctions suivantes :

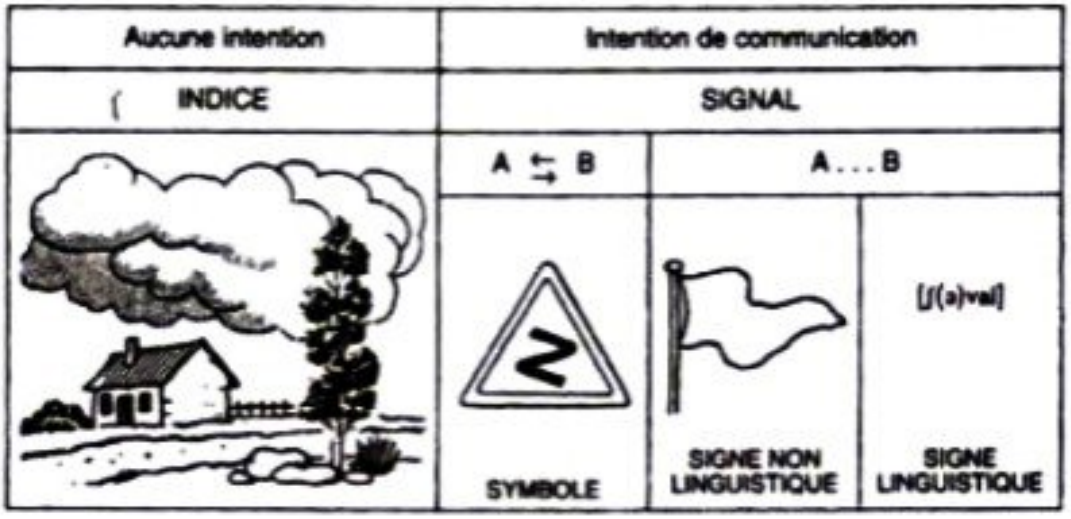
– 1) Certains signes sont produits **sans volonté de communication** et d'autres impliquent une **intention de communiquer**. Cette distinction permet déjà une première approche des signes. En effet, on peut opposer l'**indice** au **signal**. À ce propos, on donne souvent l'exemple du ciel d'orage (Georges Mounin, *Clefs pour la linguistique*, Paris, Seghers, 1987, p. 37) : le *ciel d'orage* noir et menaçant n'a pas l'intention de communiquer avec le météorologiste, mais il est cependant l'indice d'une pluie possible. La *fumée* est l'indice du feu, les *larmes* l'indice de la douleur, les *boutons sur la peau* l'indice de telle ou telle maladie, etc. L'indice peut être défini comme un fait immédiatement perceptible qui fait connaître quelque chose à propos d'un autre fait qui ne l'est pas.

Dans ces exemples, il y a un rapport physique ou d'appartenance entre l'objet représentant et la chose ou l'idée représentée. Le ciel noir est étroitement lié à la pluie. La fumée, les boutons et les larmes sont les conséquences naturelles du feu, de la maladie et de la douleur : la fumée et les boutons ne sont pas volontaires, les larmes ne sont jamais censées l'être.

Contrairement aux indices non intentionnels, il y a des signes qui impliquent une volonté de communication. Dans ce cas, on parle de **signaux**. On a vu que le *ciel d'orage* n'a pas l'intention d'annoncer le mauvais temps, mais cet indice va conduire le responsable de la sécurité d'une plage à hisser un *drapeau rouge*. Ce drapeau, qui indique que la baignade est dangereuse, est donc un fait qui a été produit artificiellement pour servir d'indice. Alors que la forme du drapeau s'offre immédiatement à la vue, la signification (= le danger) associée à cette forme doit être comprise, ce qui suppose un apprentissage préalable. Selon cette perspective, la *canne blanche* est le signal de la cécité, la *croix verte* le signal des pharmacies, le *feu vert* le signal du passage libre, le *clin d'œil* le signal de la complicité, etc.

– 2) L'intention de communiquer permet donc de faire la différence entre l'indice et le signal. L'observation des rapports qui existent entre l'objet perçu et ce qu'il représente permet une deuxième distinction : celle entre le **symbole** et le **signe**. Un *Z* sur un panneau routier annonce un virage, une *tête de cheval* indique une boucherie chevaline, un *dessin de cuiller et fourchette entrecroisées* un restaurant. Dans ces trois exemples, il y a un rapport de ressemblance formelle entre la forme de l'objet représentant et celui de l'objet représenté. Le *Z*, la *tête de cheval* et le *dessin de cuiller et fourchette entrecroisées* sont des **symboles**. Le symbole est un signal qui marque un rapport analogique, constant dans une culture donnée, avec l'élément qu'il signifie.

Cependant, il convient de souligner que la plupart du temps il n'existe aucun lien naturel entre la forme de l'objet représentant et celui de l'objet représenté. Il n'y a pas de rapport d'analogie entre un *drapeau rouge* et une baignade dangereuse, ou entre une *canne blanche* et la cécité, ou encore entre une *croix verte* et la pharmacie. Le *drapeau rouge*, la *canne blanche* et la *croix verte* sont donc des signes. Il n'y a pas, non plus, de lien d'analogie entre le signe linguistique *cheval* [ʃ(ə)val], par exemple, et l'animal désigné par ce signe. Les indices relèvent des sciences d'observation, et les signes non linguistiques et les symboles, de la sémiologie. Pour schématiser et clarifier ces notions, on peut présenter le tableau suivant :



(*Source : in* Christian Baylon et Paul Fabre, *Initiation à la linguistique*, Paris, Nathan, 1990, p. 5.)

Le **signe linguistique** est un signe particulier dans cet univers des signes, car le langage humain est un langage incomparablement plus riche, plus souple et plus efficace que n'importe quel autre langage. Comme tout système signifiant utilisé à des fins communicatives, les langues sont organisées sur deux plans :

- celui des **formes** (ou **signifiants**)
- et celui des **contenus** (ou **signifiés**).

On vient de constater que parallèlement au langage des hommes, il existe de nombreux autres systèmes de communication non linguistiques. Les systèmes de symboles ou les systèmes de signes arbitraires en font partie. La carte routière est un bon exemple des systèmes de symboles : chaque élément a sa représentation symbolique, les petits avions symbolisent des aérodromes, les petites touffes d'herbe des marais, les croix des cimetières, etc. Le code de la route, dans lequel les panneaux circulaires signifient une injonction, les panneaux rectangulaires une information, les panneaux triangulaires un danger, et ainsi de suite, forme également un système de communication non linguistique. En effet, la notion de système implique la présence de signes stables d'un message à l'autre qui se définissent fonctionnellement par leur opposition les uns aux autres.

Les langues naturelles se différencient de la plupart des autres systèmes par la propriété d'être doublement articulées. Chaque langue naturelle possède un petit nombre de **phonèmes** (= une unité de la chaîne parlée qui a une fonction différentielle, mais qui n'a pas de signification). Phonétiquement, le français possède 19 consonnes (auxquelles viennent s'ajouter deux consonnes dues à des mots empruntés à l'anglais et à l'espagnol), 3 semi-consonnes et 16 voyelles. Avec ces 38 unités sonores (*cf.* alphabet phonétique international, p. 209), on peut construire une infinité d'unités lexicales et morphologiques.

Ce type de combinaison s'appelle donc la « double articulation du langage ». On considère que les unités signifiantes constituent la première articulation, parce que c'est la couche du langage que l'on appréhende en premier. C'est elle qui véhicule le sens. Ainsi la suite phonique ou graphique : *Un enfant joue dans le jardin* se découpe en six de ces unités : *un, enfant, joue, dans, le* et *jardin*. Ces unités de première articulation sont généralement appelées **morphèmes** (= la plus petite unité ayant une signification dans la langue) pour les distinguer des **mots**, qui sont souvent constitués d'un seul morphème (ex. : *enfant, jardin, masque, juste*), mais qui peuvent aussi être formés de deux ou de plusieurs morphèmes (*enfant-in, jardin-age, dé-masqu-er, in-juste-ment* et *anti-constitution(n)-elle-ment*).

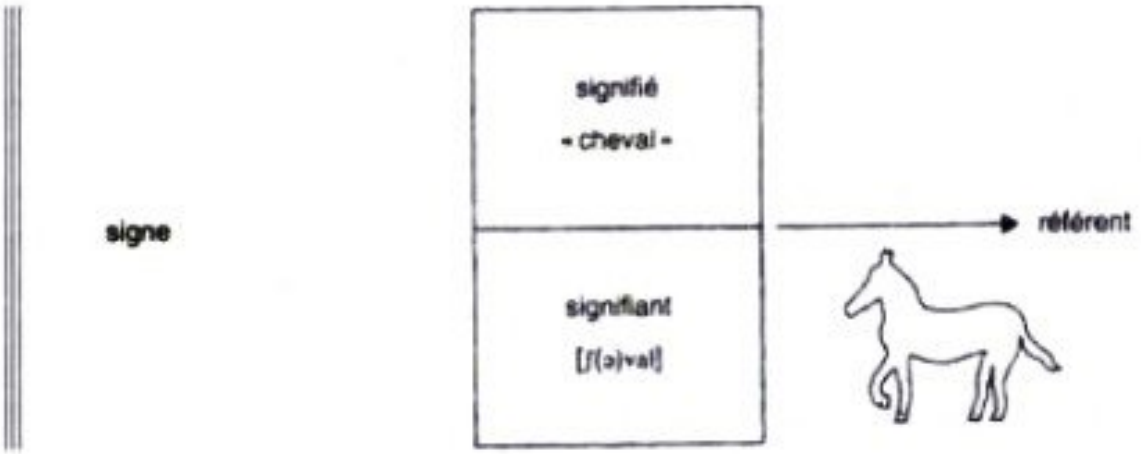
À un second niveau, les morphèmes s'articulent en segments distinctifs minimaux appelés « phonèmes ». Dépourvues en elles-mêmes de signification, ces unités de deuxième articulation ont pour unique fonction de distinguer entre elles les unités signifiantes de première articulation. Le mot *raison* [ʀɛzɔ̃], par exemple, est une combinaison de quatre phonèmes qui, comme telle, distingue ce mot des autres mots français : cette unité lexicale s'oppose en tous points à celle qui articule le mot *jardin*, mais ne se distingue que par son premier élément, r [ʀ], de celle qui articule le mot *saison* [sɛzɔ̃]. Toutes les langues naturelles sont orales avant d'être écrites, beaucoup de langues ne possèdent pas de forme écrite. Les alphabets dans les écritures alphabétiques font correspondre, bien que d'une manière souvent approximative, un nombre à peu près équivalent de lettres. Tous les énoncés d'une langue ayant adopté ce type d'écriture peuvent donc être retranscrits à l'aide d'un petit nombre de lettres. La langue française possède vingt-six lettres, quelques accents et quelques signes de ponctuation.

La double articulation donne au langage humain la créativité qui lui est propre, cette capacité d'exprimer par des combinaisons perpétuellement nouvelles des pensées constamment nouvelles.

1.2. LA NATURE DU SIGNE LINGUISTIQUE

On peut dire, en s'appuyant en partie sur des idées de Ferdinand de Saussure (*Cours de linguistique générale*, Paris, Payot, 1972), que le signe linguistique se caractérise par les traits suivants :

– a) Il est formé par l'association d'une « image acoustique », appelée « **signifiant** » et d'un concept ou « image mentale », appelé « **signifié** ». Ces deux faces du signe linguistique sont indissociables, puisque quand on prononce les sons [ʃ(ə)val], on évoque aussitôt le concept « cheval » et, inversement, ce concept ne peut exister indépendamment du matériel phonique. Ces deux faces sont donc solidaires comme le sont le recto et le verso d'une feuille de papier. Le signe linguistique réfère à un objet du monde, appelé « **référent** ».



On doit prendre garde à ne pas confondre le référent et le signifié. Le référent est un fragment de réalité et le signifié est une représentation de cette réalité. Le signifié est donc une abstraction, une espèce de réalité psychologique. Il est plus pauvre et mieux organisé que la réalité. Il simplifie la complexité du réel et met en évidence l'essentiel en donnant un premier classement des éléments du monde. Par exemple, le signifié du signe *cheval* ne tient pas compte de la diversité des chevaux qui existent, mais ne retient que ce qui est commun à tous (crinière, sabots, queue…).

– b) Le lien entre signifiant et signifié est **arbitraire** : il n'existe aucun rapport interne entre le concept, celui de « cheval » par exemple, et la suite de sons qui le représente : [ʃ] + [(ə)] + [v] + [a] + [l]. On en veut pour preuve la variété des dénominations de langue à langue pour une même réalité signifiée : français *cheval*, anglais *horse*, finnois *hevonen*, suédois *häst*. Chaque langue a sa propre façon de nommer le réel.

– c) Le rapport constitutif du signe linguistique peut être considéré comme **conventionnel**, puisqu'une fois établi, il s'impose aux usagers qui sont obligés d'accepter tels quels les signes en usage dans leur communauté. Toute infraction à la règle admise est sanctionnée socialement.

– d) Le signe linguistique donne la possibilité de parler d'objets ou de choses absents ou imaginaires. Quand on parle de chats, de démons ou de fées, on n'a pas besoin de les voir. On peut même n'en avoir jamais vu.

– e) Le signe linguistique est **typiquement humain**. Le pouvoir d'abstraction du signe linguistique fait partie des propriétés qui distinguent le langage humain du langage des animaux. D'après certains chercheurs, le gorille possède 22 cris différents, mais chacun d'eux est étroitement associé à une situation particulière. Aucun animal n'est capable de raconter ou d'évoquer un événement du passé, ou encore d'exprimer par la voix des sentiments différents en l'absence du stimulus qui aurait pu les provoquer.

La double articulation permet aussi de distinguer la communication linguistique humaine et la communication non linguistique animale. Chez certains animaux, on rencontre un système de combinatoire qui peut présenter quelques ressemblances avec la double articulation humaine. Un oiseau qui possède 5 notes de musique peut former, par exemple, 10 messages différents. Mais là s'arrête le processus. Cet oiseau ne composera jamais un 11e message.

– f) Les signes linguistiques sont la plupart du temps **arbitraires** (ou non naturels) puisque l'implication réciproque entre leurs deux faces n'est pas fondée sur une correspondance naturelle entre la forme du signifiant et les traits définitoires du signifié. D'une langue à l'autre, la même réalité notionnelle est souvent exprimée par des formes lexicales totalement différentes.

Cependant, dans chaque langue, malgré le grand nombre de signes arbitraires, on rencontre des signes qui entretiennent avec la réalité des relations moins arbitraires. Dans ce cas, on a affaire à des signes **motivés**. Un exemple de motivation du signe linguistique est offert par les **onomatopées** qui constituent une frange très marginale du lexique : *coucou, cocorico, meuh* imitent respectivement le chant d'un oiseau, du coq et le meuglement d'une vache. *Crac, boum, patatras, tic tac, froufrou, piqueter, miauler, picoter, murmurer, chuchoter, ronronner*, etc. reproduisent des bruits.

On peut pourtant constater que ces productions imitatives apparaissent elles-mêmes fortement conventionnalisées. Si la langue française a *cocorico* pour imiter le chant du coq, l'anglais a *cock-a-doodle-do*, l'italien *chichirichi* ([kiki-riki]), le japonais *kokekokko*, le hollandais *kukeleku* et le finnois *kukkokiekuu*. Cela montre que chaque langue interprète le réel selon ses propres habitudes phonologiques. Le signe linguistique, même motivé, reste donc conventionnel à l'intérieur d'une même communauté. Les onomatopées sont rarement compréhensibles aux membres d'une autre collectivité linguistique.

La motivation du signe linguistique s'observe aussi dans les formes complexes ou construites. Les signes du lexique français peuvent parfois apparaître motivés du point de vue morphologique : le *pommier* est l'arbre dont le fruit est la pomme, l'*abricotier* produit des abricots, le *cerisier* des cerises, le *citronnier* des citrons, le *châtaignier* des châtaignes, etc. De même, l'*ourson* est le petit de l'ours et le *chaton* le petit du chat. Dans ces exemples, la motivation résulte de l'emploi des procédés de dérivation.

Les mots dérivés, contrairement à leurs bases totalement arbitraires, sont relativement motivés. *Pommier* a été formé à partir de *pomme* à l'aide du suffixe *-ier. Abricotier, cerisier, citronnier, châtaignier*, ainsi qu'un grand nombre d'autres noms d'arbres fruitiers ont été créés de la même façon. Un signe à motivation relative est donc nécessairement complexe.

La motivation relative se rencontre aussi dans les mots composés combinant des signes élémentaires immotivés. Contrairement à *cent*, la forme composée *quatre-vingt-dix-neuf* s'interprète analytiquement comme le résultat de la multiplication de *vingt* par *quatre* auquel s'additionne la somme de *dix* et de *neuf*.

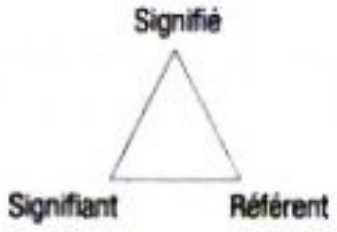
La poésie exploite les rapports qui existent entre le niveau phonique et le niveau sémantique du langage. Autrement dit, elle cherche à mettre les récurrences phoniques non signifiantes en relation avec le sens et ainsi à rendre le lien qui existe entre le signifiant et le signifié du signe linguistique moins arbitraire, plus naturel. Tout le monde sait que les poètes favorisent les onomatopées, les créations lexicales, les changements de sens, etc.

– g) Qu'il s'agisse de leur structure interne ou de leurs combinaisons, les signes linguistiques sont **linéaires**. Ce caractère linéaire du signe est dû à la nature orale du langage. Il est tout à fait impossible de prononcer simultanément deux sons, deux syllabes ou deux mots. La manifestation orale du langage se déroule donc dans le temps. Cette linéarité se répercute sur la transcription alphabétique qui se déroule dans l'espace : on ne peut pas écrire les unes sur les autres les différentes unités graphiques de la langue. Les lettres et les mots se succèdent sur la dimension de la ligne.

Le langage exploite doublement cette dimension unique. D'une part, quand on emploie les trois phonèmes */p/, /a/* et */l/*, leur ordre dans */pal/* « pal », */alp/* « alpe » et */pla/* « plat » est extrêmement important pour la signification du message. D'autre part, quand on dit : *Pierre bat Paul*, cela ne veut pas dire la même chose que *Paul bat Pierre*. Cette caractéristique différencie les langues naturelles humaines de tous les systèmes de communication qui construisent leurs messages sur la trame de l'espace et non sur celle du temps.

1.3. LE SIGNE LINGUISTIQUE ET LE RÉFÉRENT

Le signe linguistique est souvent présenté sous forme d'un triangle.



Les signes linguistiques ont la propriété de pouvoir renvoyer aux objets du monde, extérieurs à la langue ; ces objets sont des **référents** qui peuvent être des entités matérielles ou conceptuelles (objets, êtres, processus, lieux, événements, etc.). Ils relèvent de l'univers extralinguistique réel ou fictif (par exemple, la licorne ou le dragon). Il est important de souligner que le processus référentiel de mise en relation d'unités linguistiques et d'unités extralinguistiques n'est pas unitaire. Selon Catherine Kerbrat-Orecchioni (*L'Énonciation de la subjectivité dans le langage*, Paris, A. Colin, coll. « Linguistique », 1997, p. 35), on peut faire la différence entre une référence absolue, une référence relative au contexte linguistique et une référence relative à une situation d'énonciation.

Dans le cas de la **référence absolue**, l'opération de référenciation se caractérise par son autonomie. Cela veut dire que le repérage du référent dépend uniquement des informations contenues dans l'expression référentielle. Par exemple, la construction référentielle de l'énoncé *la rose est une fleur*, ne dépend que de la connaissance des mots et des savoirs extralinguistiques des interlocuteurs.

Dans le cas de la **référence relative**, les expressions référentielles ne disposent pas d'un contenu descriptif suffisant pour permettre un acte de référenciation autonome. La référence doit se réaliser grâce à des informations différentes de celles qui sont disponibles dans les unités linguistiques. Si les données convoquées font partie de l'environnement linguistque de l'unité référentielle, la référence est dite « **relative au contexte linguistique** » ou « **anaphorique** ». Par exemple, dans l'énoncé *Marine est une fille intelligente, mais elle ne travaille pas assez*, l'interprétation référentielle du pronom *elle* dépend de la prise en compte d'un élément du contexte linguistique, le nom propre *Marine*. Cette procédure référentielle anaphorique est indirecte dans la mesure où elle effectue un détour par le contexte linguistique pour désigner le référent visé.

Si les données convoquées coïncident avec l'un des éléments constitutifs de la situation d'énonciation, la référence est « **relative à la situation d'énonciation** » ou « **déictique** ». Par exemple, si l'on dit *Est-ce que tu aimes ça ?*, l'objet désigné par le pronom démonstratif *ça* est repérable dans la situation de communication. Cette procédure référentielle est directe, elle n'est pas médiatisée par une autre forme linguistique présente dans l'énoncé.

1.4. LES RELATIONS SYNTAGMATIQUES ET PARADIGMATIQUES

Les mots entretiennent entre eux deux types de relations fondamentales : les relations syntagmatiques qui s'observent entre les termes d'une même construction et les relations paradigmatiques qu'on peut établir entre une unité et toutes celles qui pourraient la remplacer dans un environnement donné.

Ainsi dans la phrase : *La petite fille joue avec son chat*, l'adjectif *petite* est en **relation syntagmatique** avec l'article défini *la* qui le précède et avec le substantif *fille* qui le suit. Ce groupe nominal est en relation syntagmatique avec le verbe *joue*, et ainsi de suite. Les relations syntagmatiques affectent donc des éléments qui sont présents dans l'énoncé. On leur donne souvent le nom de relations *in praesentia* (lat. : « en présence »).

Dans la même phrase, l'article défini *la* est en **relation paradigmatique** avec d'autres déterminants : *cette, ma, une, notre*, etc. ; l'adjectif *petite* avec d'autres adjectifs tels que *grande, jolie*, etc. ; le substantif *fille* avec des substantifs comme *voisine, cousine, dame*, etc. ; le verbe *joue* avec d'autres verbes tels que *court, danse*, etc. Les éléments en relation paradigmatique sont mutuellement substituables dans un environnement donné, s'y excluent les uns les autres et forment ensemble un paradigme. Les déterminants *la, cette, ma, une, notre*, etc. font partie du même paradigme, tout comme les adjectifs *petite, grande, jolie*, etc. Le paradigme est défini comme un ensemble d'unités virtuellement substituables dans un contexte donné. Les relations paradigmatiques existent entre des termes qui ne sont pas présents dans l'énoncé. On les appelle souvent relations *in absentia* (lat. : « en absence »). Elles s'opposent par là aux relations syntagmatiques.

2. LE MOT ET LE MORPHÈME

2.1. LE MOT

Le lexique désigne conventionnellement l'ensemble des mots au moyen desquels les membres d'une communauté linguistique communiquent entre eux. Cette définition du lexique oblige à donner une définition précise de l'**unité lexicale**, du **mot** en tant qu'élément de base de l'ensemble. La notion traditionnelle de mot est l'une de celles qui ont tendance à solliciter le plus constamment l'attention des linguistes. Cette notion, qui semble être familière et évidente pour le grand public, constitue pour le linguiste une source de difficultés théoriques considérable. Dans une langue telle que le français, on arrive à donner une définition simple et rigoureuse du mot uniquement au niveau de la manifestation graphique, où le mot est un groupement de lettres, séparé, à gauche et à droite, par un blanc des autres éléments du texte. Mais on constate très vite que ces segments ne correspondent pas à une réalité linguistique bien déterminée. Les blancs du texte ne s'harmonisent que rarement avec les limites du code oral. On peut malgré tout chercher à mettre en lumière les **critères phoniques** qui aident à identifier et à délimiter les unités lexicales dans la chaîne parlée.

2.1.1. POINT DE VUE PHONIQUE

L'accent, les contraintes phonologiques et le coup de glotte démarcatif peuvent contribuer à l'identification des unités lexicales du point de vue sonore. Dans un grand

nombre de langues, l'**accent** joue un rôle démarcatif important. C'est le cas des langues à **accent lexical**. Dans certaines langues (ex. : tchèque, finnois, hongrois), l'accent tombe sur la première syllabe du mot ; dans d'autres, il peut occuper diverses places, mais toujours la même pour un mot donné (ex. : anglais, italien, russe). Le français est une langue à **accent syntaxique**. Si le mot est isolé, l'accent se place toujours sur la dernière voyelle du mot à l'exception du *e* caduc. Mais le mot ne fonctionne que rarement seul : la plupart du temps, il est inséré à l'intérieur d'un syntagme, d'un groupe et dans ce cas, l'unité d'accentuation est beaucoup moins le mot que le syntagme.

maisón
une maison rouge

En outre, en français, en dehors de cet accent dit « interne », non emphatique, qui est purement linguistique, il existe un accent dit « externe » ou emphatique qui est lié à des facteurs expressifs et intellectuels. En effet, l'accent peut être déplacé sous l'effet d'une volonté d'expressivité : l'accent du mot *formidable !*, qui frappe normalement la voyelle */a/* de la dernière syllabe peut très bien passer sur le */o/* initial. En français, le rôle démarcatif de l'accent dans l'identification des mots est donc très faible.

Les **contraintes phonologiques** peuvent également aider l'auditeur à identifier et à délimiter les unités lexicales dans la chaîne sonore. Il y a des phonèmes et des groupes de phonèmes qui sont possibles à toutes les places du mot. Dans ce cas, leur valeur démarcative est nulle. En revanche, certains phonèmes ou combinaisons de phonèmes sont impossibles à telle ou telle place du mot et peuvent ainsi donner des indications sur les limites initiales ou finales de l'unité lexicale. Alors que la présence d'un /ø/, possible à toutes les places (ex. *heureux, heureusement* : [øʀø], [øʀøzmɑ̃]) n'enseigne rien, celle d'un /œ/, impossible à la finale absolue, signale qu'on n'est pas arrivé à la fin d'un mot (*cf.* Jacqueline Picoche, *Précis de lexicologie française*, Paris, Nathan, 1992, p. 14).

Il est intéressant de noter que le mot français isolé peut commencer par n'importe quelle consonne. Toutefois, les mots commençant par les semi-consonnes [j, w, ɥ], ainsi que ceux qui ont pour initiale un *h* dit « *aspiré* » sont très rares :

ouaille, ouate, ouest, ouïe, ouistiti, etc.
hachis, haïr, haine, hardiesse, etc.

Ce que l'on appelle aujourd'hui un *h aspiré* en français, n'est en fait ni un [h], ni aspiré. Il s'agit d'un phénomène d'absence de liaison (*les hachis* : on ne dit pas [lezaʃi]*, mais [le'aʃi]) et d'absence d'élision (on ne dit pas *l'hachis** [laʃi], mais *le hachis* [lə'aʃi]). Le *h* dit « *muet* » permet l'élision et la liaison, comme dans *l'homme* [lɔm] et *les hommes* [lezɔm].

Les trois consonnes nasales [m], [n] et [ɲ] peuvent toutes apparaître à l'intermédiaire : *hameau, anneau, agneau*, [amo], [ano], [aɲo] et à la finale des mots : *came,*

canne, khâgne [kam], [kan], [kaɲ]. Mais, à l'initiale, alors que les mots commençant par [m] ou [n] sont très nombreux, les mots commençant par [ɲ], comme *gnocchi* [ɲɔki], *gnôle* [ɲol] ou *gnognote* [ɲɔɲɔt], se comptent sur les doigts d'une main. En revanche, le *h* aspiré ne s'entend jamais dans une autre position qu'au début du mot. D'une façon générale, les consonnes (occlusives et constrictives) ont du mal à se maintenir à la finale du mot en français : par exemple, la consonne qui apparaît dans la flexion [simât+e] (*cimenter*) ou tombe quand le radical est nu [simâ] (*ciment*).

Certaines langues (ex. l'allemand) connaissent le **coup de glotte démarcatif** qui a pour but de signaler la frontière des mots composant un syntagme. Ce phénomène est très rare en français (*le un, le onze, le harnais, un haricot* : [lə'œ̃], [lə'ɔ̃z], [lə'ʁnɛ], [œ̃'ʁiko]) qui a tendance à favoriser toutes sortes de liaisons et d'élisions.

Si l'on essaie de trouver des critères phoniques de délimitation du mot, on risque d'être déçu. Les blancs du texte écrit ne correspondent que rarement aux limites du code oral. On vient de constater qu'en français le rôle démarcatif de l'accent dans l'identification des mots est très faible par rapport aux accents des langues à accent lexical. Les locuteurs peuvent souvent faire entendre des accents ou effectuer des pauses en des points différents d'une même séquence sonore. On a également vu que les contraintes phonologiques ne peuvent aider l'auditeur que partiellement et que les liaisons et les élisions de toutes sortes ont tendance à contribuer à effacer les frontières entre les différents mots composant un syntagme.

Les critères sonores ne permettent donc pas aux linguistes d'appréhender le fonctionnement réel de l'élément mot. En effet, les traits démarcatifs phoniques ne semblent jouer, dans le découpage en mots de la chaîne parlée, qu'un rôle secondaire, les critères principaux étant d'ordre syntaxique et sémantique.

2.1.2. POINT DE VUE SYNTACTICO-SÉMANTIQUE

Dans les paragraphes suivants, on cherchera à mettre en relief les critères qui aident à identifier et à délimiter les unités lexicales dans la chaîne parlée en se plaçant sur le plan syntactico-sémantique.

Les unités lexicales sont inscrites comme unités dans le code de la mémoire du sujet parlant qui doit les reproduire telles quelles, en bloc, dans le discours. Le locuteur n'a pas la liberté de changer, à sens égal, l'ordre des éléments dans ces unités (*maisonnette, *ettemaison, pomme de terre, *de terre pomme*), ni de les séparer par des insertions (**pomme jaune de terre*).

En effet, sur le plan syntaxique on peut isoler des unités de fonctionnement en utilisant divers critères dont les principaux sont les procédés d'**inséparabilité** et de **commutation**. Ces manipulations ont pour but de segmenter les unités de la chaîne parlée et de les analyser (*cf.* Joëlle Gardes, *La Grammaire. Méthodes et notions*, « Cursus », Paris, Armand Colin, 2012).

Le critère d'inséparabilité, c'est-à-dire l'impossibilité d'insérer un élément quelconque à l'intérieur d'une unité lexicale, fonctionne bien évidemment pour les unités morphologiquement et graphiquement simples telles que *fille, maison, ministre, moustique*, ainsi que pour les unités morphologiquement composées, mais graphiquement simples, telles que *in-perturb-able, ir-récupér-able, em-poisonn-eur ;* mais son rôle essentiel est de tester le caractère lexical d'unités graphiquement complexes telles que *pomme de terre* ou *assistante sociale*. Il est tout à fait impossible de dire **pomme jaune de terre* ou **assistante très sociale*. Cela prouve que le degré de cohésion des unités lexicales est très fort.

La commutation est une opération linguistique qui permet de remplacer un élément par un autre :

La *pomme de terre* est un légume qui contient des vitamines.
Le *navet* --
La *carotte* --

Le *chemin de fer* a coûté cher à l'État.
La *route* --

L'*assistante sociale* s'occupe bien des enfants maltraités.
L'*infirmière* --
Le *médecin* --

Le critère de commutation, qui permet de mettre en évidence des classes d'équivalences (*pomme de terre, navet, carotte*), confirme ce que révèle le critère d'inséparabilité : lorsqu'une unité complexe est lexicalisée (*pomme de terre, assistante sociale, chemin de fer*), c'est-à-dire sentie comme un mot, elle joue exactement le même rôle qu'une unité simple (*navet, carotte ; route ; infirmière, médecin*).

Il faut signaler qu'à l'aide de ces procédures, on peut isoler des unités lexicales très diverses dans la chaîne parlée. La notion de mot regroupe donc des éléments différents que la langue fournit au locuteur pour construire des énoncés :

la, son, notre, puis…
fille, maison, garçon, moustique, ministre, table, chaise, timide…
fillette, maisonnette, anticonstitutionnellement, inséparable, timidité, intimidé…
donnerons, chantera, dansais, courir…
bouilleur de cru, chemin de fer, pomme de terre, machine à coudre…
qu'en dira-t-on, à pas de loup, au fur et à mesure…

Toutes ces unités sont en quelque sorte des unités préfabriquées, stockées dans la mémoire lexicale du sujet parlant qui les combine pour former des phrases selon les

besoins de la communication. Le mot est donc avant tout une structure phonique et graphique stable, que l'on apprend à reconnaître et à reproduire.

2.2. LES MOTS SIMPLES ET LES MOTS CONSTRUITS

La **morphologie** (du grec *morphê* « aspect, forme ») est la discipline linguistique qui étudie les **morphèmes** (les plus petites unités de signification de la langue). Le **mot**, c'est-à-dire la plus petite forme linguistique ayant une autonomie, peut contenir un ou plusieurs morphèmes.

2.2.1. LES MOTS SIMPLES

Quand le mot est formé d'un seul morphème, il s'agit, d'après une terminologie assez courante, d'un **mot simple** ou (**mot monomorphématique**) :

fille, maison, ministre, moustique, timide, garçon, femme…

Le morphème est défini comme la plus petite unité de signification de la langue. Par exemple, *fille* [fij] est un morphème qui est segmentable en **phonèmes** */f/-/i/-/j/*, unités qui de par leur combinaison contribuent à la signification mais ne sont pas, en elles-mêmes, porteuses d'un sens. Il est important de noter qu'un mot monomorphématique peut comporter une ou plusieurs **syllabes** :

fille [fij], femme [fam]
timide [ti-mid], moustique [mus-tik], garçon [gaʁ-sɔ̃]

Il y a des morphèmes qui ne contiennent qu'un seul phonème. C'est le cas, par exemple, de la préposition *à* [a], du pronom *y* [i] ou de l'article élidé *l'* [l], mais, puisque ces unités ont un sens, elles ont un statut différent du phonème.

On constate que le mot-morphème est une unité à cheval sur deux niveaux de codification et c'est pour cela qu'il est une source permanente de confusion. Le morphème constitue bien une unité linguistique, parce qu'il existe des procédures pour l'isoler. Mais il n'en va pas de même du **mot**, qui peut être défini comme la forme linguistique la plus petite qui ait une autonomie. Le phonème, en effet, n'en a aucune et il en va de même de bon nombre de morphèmes qui ne se rencontrent jamais à l'état libre et ont besoin d'entrer en combinaison avec un autre morphème : ainsi en est-il du morphème *-eur* que l'on trouve dans *coiffeur*, du morphème *re-* que l'on trouve dans *refaire*, etc. Un morphème autonome acquiert donc le statut de mot, de même qu'un phonème qui présente un sens constitue un morphème.

2.2.2. LES MOTS DÉRIVÉS ET LES MOTS FLÉCHIS

Une très grande partie des mots (environ 75 %) du lexique de la langue française sont composés de deux ou de plusieurs morphèmes. Dans ce cas, on a souvent

affaire à des **mots dérivés** (type *danseur*) et à des **mots fléchis** (type *danserons*). La **commutation** est une opération linguistique qui permet de segmenter les mots en morphèmes. *Danseur*, par exemple, s'analyse en deux morphèmes [dɑ̃s] - [œʀ], *danserons* s'analyse en [dɑ̃s] - [(ə)ʀ] - [ɔ̃], chacun des segments ainsi dégagés étant porteur d'un sens. La condition nécessaire, en effet, pour qu'une partie de mot puisse constituer un morphème est qu'elle puisse être remplacée par un autre élément, donc commuter avec lui :

dans [dɑ̃s] -eur [œʀ]
-euse [øz]
-er [e]
etc.

La commutation doit obligatoirement être pratiquée sur les deux parties du mot :

dans- [dɑ̃s] -eur [œʀ]
march- [maʀʃ]
chant- [ʃɑ̃t]
camp- [kɑ̃p]
etc.

Il est donc nécessaire que les éléments ainsi isolés présentent un sens. Le morphème *-eur* dans *danseur, chanteur, marcheur, campeur, etc.* désigne l'agent de l'action, celui qui danse, chante, marche, etc. Si une forme se trouve associée à des sens très différents, on posera des morphèmes différents. Dans *blancheur, fraîcheur, grandeur, largeur*, etc., la commutation permet également d'isoler une forme *-eur*, mais ce morphème, au lieu de signifier l'agent de l'action, désigne la qualité de ce qui est blanc, frais, grand, etc. On est donc en présence de deux **morphèmes homonymes** (= deux formes de prononciation identique, mais de sens différent).

• Les différents types de morphèmes

La commutation aide à diviser les morphèmes en deux grandes classes :
- 1) les morphèmes lexicaux ou lexèmes ;
- 2) les morphèmes grammaticaux ou grammèmes.

Les **morphèmes lexicaux** permettent au mot d'avoir une autonomie sémantique, tandis que les **morphèmes grammaticaux** insèrent le mot dans des séries et indiquent souvent ses relations avec d'autres éléments de la phrase. Ainsi, dans le mot dérivé *danseur, dans-* est un morphème lexical qui permet de distinguer le mot des autres mots de la même série : *chanteur, marcheur, voleur, campeur*, etc., tandis que *-eur* est un morphème grammatical qui indique l'agent de l'action. Le mot fléchi *danserons* se décompose en

dans- [dɑ̃s], morphème lexical exprimant l'idée de « danser », *-er-* [(ɔ)ʀ], morphème grammatical exprimant l'idée du futur, et *– ons* [ɔ̃], autre morphème grammatical, exprimant l'idée de la première personne du pluriel.

Ces deux types d'unités se distinguent par les faits suivants : les morphèmes lexicaux sont extrêmement nombreux et leur liste est ouverte dans la mesure où elle accueille régulièrement de nouvelles unités. En ce qui concerne les morphèmes grammaticaux, ils sont en nombre restreint et leur liste est fermée. On arrive à les répertorier.

• Les morphèmes grammaticaux

Les morphèmes grammaticaux jouent un rôle décisif dans l'organisation grammaticale de la phrase, qu'il s'agisse des **marques morphosyntaxiques** (nombre, genre, personne, temps et mode) ou des **mots-outils** qui marquent les relations entre mots et groupes de mots dans la structure phrastique (prépositions et conjonctions) ou qui assurent l'actualisation d'une autre partie du discours (les déterminants). Les morphèmes grammaticaux servent à transmettre des notions générales souvent axées sur la situation d'énonciation, contrairement aux morphèmes lexicaux qui véhiculent des concepts ayant leur spécificité propre.

Il est important de signaler que parmi les morphèmes grammaticaux, on distingue les **morphèmes liés** ou **affixes** (*-eur, -ité, -er, -ons*), qui ne peuvent apparaître que dans le cadre de l'unité mot, et les **morphèmes non liés**, comme *le* (article ou pronom), *il* (pronom de 3e personne), etc., auxquels on accorde généralement le statut de mot.

• Les affixes dérivationnels et les affixes flexionnels

Parmi les morphèmes grammaticaux liés ou affixes, on peut distinguer, à leur tour deux types de morphèmes : les **affixes dérivationnels** (*-eur, -ité*) et les **affixes flexionnels** (*-ons, -ez, -er-*...). Les affixes dérivationnels ont essentiellement une fonction sémantique. Ils servent à créer une nouvelle unité lexicale à partir d'un mot déjà existant :

chant (er) (verbe) → chant-*eur* (nom)
passif, -ive (adjectif) → passiv-*ité* (nom)
légal (adjectif) → *il*-légal (adjectif)
loyal (adjectif) → *dé*-loyal (adjectif)

Les affixes qui se placent devant la base sont appelés « **préfixes** » (*il-légal, déloyal*), et les affixes qui se trouvent après la base se nomment « **suffixes** » (*chant-eur, passiv-ité*).

Les affixes flexionnels ne créent jamais de nouvelles unités lexicales. Au contraire, ils indiquent les rapports que la base entretient avec l'énoncé où elle est employée : par exemple, le morphème de nombre dans un adjectif indique avec quel substantif cet adjectif

est lié : *petits enfants*. La morphologie flexionnelle comprend la **flexion nominale**, c'est-à-dire la variation de forme (le genre et le nombre) du substantif et de l'adjectif :

cheval, chevaux
chat, chatte
chien, chienne
grand, grande
petit, petite
gros, grosse

et la **flexion verbale** qui s'occupe des marques des temps, des personnes et des modes des verbes :

chant-er-ons (*-er-* : futur, *-ons* : 1re personne du pluriel)
chant-i-ez (*-i-* : imparfait, *-ez* : 2e personne du pluriel)

Un affixe flexionnel ne modifie jamais la catégorie grammaticale de la base à laquelle il s'adjoint. Si la base est un verbe, l'élément obtenu après l'adjonction du suffixe reste un verbe. Au contraire, un affixe dérivationnel peut changer la catégorie de sa base, même si cela ne se produit pas dans tous les cas.

En ce qui concerne la régularité, on peut constater qu'un affixe flexionnel entre dans une série close, comme celle des terminaisons verbales, et qu'il se combine avec toutes les bases d'un même type : toutes les bases verbales se combinent avec les affixes de la conjugaison. Au contraire, l'adjonction d'un affixe dérivationnel est beaucoup moins régulière. La langue française possède les dérivés tels que *campeur, marcheur, coiffeur, danseur*, etc., mais **souffreur*, **aimeur* ou **détesteur* n'existent pas.

Il est indispensable de faire la différence entre la **base** et le **radical** d'un mot dérivé. Si l'on retire un affixe à un mot, on obtient la base sur laquelle il est construit : la base à laquelle s'adjoint le suffixe *-ment-* dans *affermissement* est *affermisse* (base longue du verbe *affermir*). La base à laquelle s'adjoignent le préfixe *a-* et le suffixe *-ir* dans ce dernier mot est *ferme*. La base est donc un élément qui peut contenir plusieurs morphèmes. Lorsque tous les affixes ont été enlevés, il reste une base minimale que l'on appelle « radical ».

2.2.3. LES MOTS COMPOSÉS ET LES LOCUTIONS

Parmi les unités lexicales de la langue française, on peut également rencontrer des unités qui sont souvent (pas toujours) graphiquement complexes, mais qui fonctionnent exactement comme les mots graphiquement simples :

pomme de terre
chou-fleur

chemin de fer
machine à coudre
anthropologue
pourboire
qu'en dira-t-on
à pas de loup

À l'intérieur de ces unités, on peut distinguer les **mots composés** et les **locutions**. La **composition** peut être définie comme la juxtaposition de deux éléments qui peuvent servir de base à des dérivés. Certains linguistes ont tendance à la définir comme la juxtaposition de deux éléments qui peuvent exister à l'état libre. Cette définition permettrait bien d'y intégrer des mots comme *chou-fleur* ou *porte-mine*, mais pas des mots comme *anthropologue* où ni *anthrope* ni *logue* ne se rencontrent seuls, alors même qu'à la différence des suffixes et des préfixes ils peuvent servir de base à des dérivés : *anthropien, logistique*, etc. On ne peut donc voir en eux des affixes. Ce sont des bases dont le fonctionnement est fondamentalement le même que celui de *fleur* ou de *porte*.

Très souvent on classe parmi les mots composés les unités complexes, principalement nominales, à deux termes (*chou-fleur, porte-mine…*), et les unités complexes, uniquement nominales, à trois termes (*pomme de terre, chemin de fer, machine à coudre…*). Tous les faits de résultats différents et de formes plus complexes sont regroupés sous l'étiquette « locutions » (*à pas de loup, qu'en dira-t-on, au fur et à mesure…*).

Les mots composés et les locutions sont donc inscrits, à l'instar des mots monomorphématiques, des mots dérivés et des mots fléchis, comme unités figées dans le code de la mémoire du sujet parlant. Cela veut dire que l'on doit les reproduire tels quels, en bloc, dans le discours. Au contraire, le syntagme libre et la phrase sont des unités qui ne sont pas codées. Leurs éléments peuvent être choisis, déplacés et intervertis à volonté, dans les limites de la grammaticalité et de la sémanticité.

Toutes ces différentes sortes d'unités lexicales seront étudiées en détail dans la partie consacrée à l'analyse morphologique du lexique (p. 46).

2.2.4. LES DIFFÉRENTES PARTIES DU DISCOURS

Le lexique se répartit en classes fonctionnelles traditionnellement appelées « parties du discours ». Elles concernent la **nature** du mot, laquelle s'oppose à sa **fonction** : *chat* fait partie de la classe des noms ; il a la fonction de sujet dans *le chat ronronne*. En général, la classification grammaticale distingue huit catégories différentes :

- 1) verbes (*mentir, chanter, aimer, courir…*) ;
- 2) noms ou substantifs (*chat, maison, table, amour, haine…*) ;
- 3) déterminants (*le, une, ce, mon, nos, deux, plusieurs…*) ;

- 4) adjectifs (*beau, riche, intelligent, petit, long...*);
- 5) pronoms (*je, il, celui, ceux, le mien, la mienne, nulle...*);
- 6) adverbes (*lentement, vite, peu, beaucoup, vaguement...*);
- 7) prépositions (*à, de, vers, pour, depuis...*);
- 8) conjonctions (*et, mais, quand, parce que...*).

Chaque classe peut, à son tour, se subdiviser en sous-classes (ex. : déterminants articles, dét. démonstratifs, dét. possessifs, dét. indéfinis et cardinaux, etc.). D'après certaines grammaires, les interjections (*ah!, oh!, eh!...*) forment la 9e catégorie grammaticale. Cette classification est valable pour de très nombreuses langues.

Une partie du discours est une **classe d'équivalences** qui rassemble les mots en catégories fonctionnant de façon identique. Les règles grammaticales s'appliquent ensuite de façon générale à ces classes. La classification des mots en parties du discours s'établit selon des critères morphologiques, sémantiques et syntaxiques. Le substantif et le verbe, par exemple, s'opposent déjà par les morphèmes flexionnels qu'ils mettent en jeu (genre et nombre pour le premier, personne, temps et mode pour le second). Ces deux classes s'opposent aussi par leur fonctionnement syntaxique. Le substantif est régi par le verbe, alors que le verbe ne l'est par rien. Le substantif et le verbe sont différents aussi du point de vue sémantique. Le premier renvoie à un être, un individu ou une chose, alors que le verbe dénote une action, un événement ou un état.

Dans l'ensemble des mots se manifeste la même dichotomie que dans l'ensemble des morphèmes; il existe des **mots grammaticaux** et des **mots lexicaux**. En général, les noms, les verbes, les adjectifs et les adverbes font partie des mots lexicaux. Les pronoms, les déterminants, les prépositions et les conjonctions sont regroupés parmi les mots grammaticaux.

Les mots lexicaux forment de très loin le sous-ensemble le plus important. En français, les mots grammaticaux ne dépassent guère la centaine, bien qu'ils composent à eux seuls 50 % du discours.

3. LE LEXIQUE ET LA GRAMMAIRE

On vient de voir que le terme de lexique peut recouvrir un certain nombre de notions qui sont dépendantes des critères adoptés dans la description linguistique. Il est important de signaler aussi que le **lexique** est très couramment opposé à la **grammaire**. Selon cette perspective, il s'agit en quelque sorte de deux domaines complémentaires : la grammaire fournit les règles qui permettent de combiner les mots et les groupes de mots pour former des phrases et le lexique représente l'ensemble des unités qui constituent son matériau de base.

Cependant, cette distinction n'est pas toujours aussi simple que cela : la frontière tracée entre ces deux domaines complémentaires a tendance à être fragile et floue.

Lexique et grammaire entretiennent en effet des relations fort complexes. On peut noter que les grammaires répertorient des listes de mots (les prépositions, les pronoms, par ex.) dont elles s'attachent à décrire le fonctionnement, et que dans les dictionnaires, on rencontre couramment des indications sur le genre des mots, leur accord et leur place dans la phrase, des articles comme prépositions, adverbes, etc., alors que tous ces points devraient faire l'objet des grammaires.

En outre, dans les unités lexicales comme *blanche, garçons, chanterons*, une partie au moins appartient au domaine de la grammaire ; de même, dans des mots comme *étrangement* et *diminution*, la grammaire reconnaîtra un segment final qui est à l'origine d'un changement de catégorie. On a vu dans les paragraphes précédents que les morphèmes se divisent en deux grandes classes : morphèmes lexicaux et morphèmes grammaticaux. Il va de soi que la description des morphèmes grammaticaux relève essentiellement de la grammaire et celle des morphèmes lexicaux du lexique.

On constate cependant rapidement que les mots répertoriés dans une grammaire forment une infime partie du lexique et que les règles de grammaire ne sont pas toutes explicitées dans le dictionnaire. D'autre part, s'il s'agit d'opposer des règles aux unités auxquelles elles s'appliquent, le dictionnaire présente généralement de graves lacunes ; on peut s'étonner de ne pas y trouver la liste complète des unités signifiantes que sont les morphèmes liés. De plus, bien qu'il soit impossible de parler une langue sans employer de noms propres, ceux-ci ne sont pas répertoriés dans le dictionnaire de langue, mais seulement, lorsqu'ils sont notoires, dans l'encyclopédie.

Soulignons aussi que la plupart des usagers d'une langue parviennent à maîtriser la grammaire de leur propre langue dont les règles sont en nombre limité. Ils n'ont aucun mal à reconnaître une phrase incorrecte d'une phrase correcte. Un grammairien professionnel peut atteindre une **compétence grammaticale** optimale. En revanche, aucun usager n'est capable de maîtriser le lexique de sa langue, parce que celui-ci est formé d'un nombre d'unités incalculable. Aucun lexicologue ou lexicographe ne peut espérer acquérir une **compétence lexicale** optimale. Du point de vue de l'apprentissage de la langue, on constate que les connaissances grammaticales sont acquises assez tôt une fois pour toutes, alors que les connaissances lexicales ne cessent de s'enrichir au cours de la vie de l'usager.

Les chercheurs ne disposent d'aucune procédure valable pour recenser le lexique d'une langue à un moment donné. D'après les estimations, le lexique d'une langue de civilisation comme le français ou l'anglais dépasse 200 000 mots en excluant les noms propres. On pourrait probablement atteindre facilement le chiffre de 500 000 mots avec les nomenclatures terminologiques. Or, un usager moyennement cultivé se sert peut-être uniquement d'un dixième du lexique global.